

To Extend or Not to Extend?
Context-driven Corpus Enrichment
Data Linking for Humanities Research – Workshop

Tanya Braun

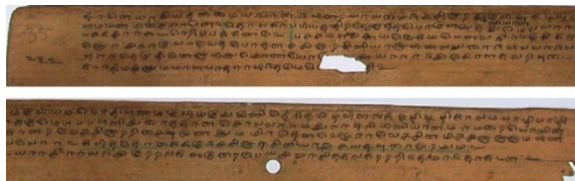
Institute of Information Systems
University of Lübeck

October 21, 2019

A sneak preview of a paper of the same name
by Kuhr, B, Bender, and Möller
accepted for publication at AI 2019

Understanding Written Artefacts

Eva Wilden, Tamil Satellite Stanzas: Genres and Distribution



Transcript

*pāra+ tolkāppiyamum pattuppāṭṭum kaliyum
āra+ kuṟuntokaiyu! aiññāṅkum – cāra+
tīru+ taku mā muṇi cey cintāmaṇiyum
virutti naccīṇārkkīṇiyamē.*

Translation

On the weighty *Tolkāppiyam* and the *Pattuppāṭṭu* and *Kali*
and on five [times] four in the ornamental *Kuṟuntokai* and on the
essential *Cintāmaṇi* made by the brilliant great sage (Tirutakkāṭēvar)
[are] the elaborate commentaries [attributed] to Naccīṇārkkīṇiyar.

Annotations for transcript or translation

- Hyperlinks (doi)
- Tags (strings)
- Descriptions of objects (entities) and relations between them

Annotations

Descriptions of objects (entities) and relations between them

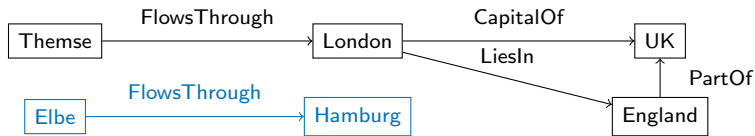


Figure: Objects and their relations in a graph

Annotations

Descriptions of objects (entities) and relations between them

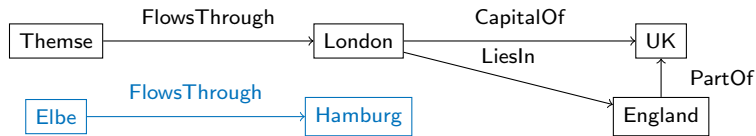


Figure: Objects and their relations in a graph

Long-term Goal

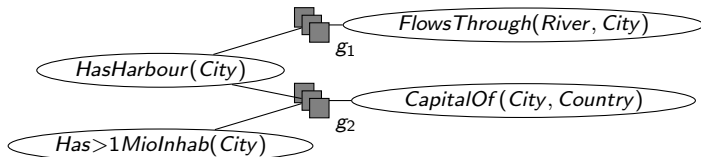
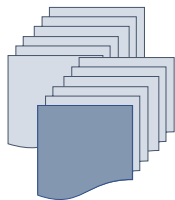


Figure: A probabilistic relational model allowing for modelling uncertainty

A Corpus of Documents and Annotations

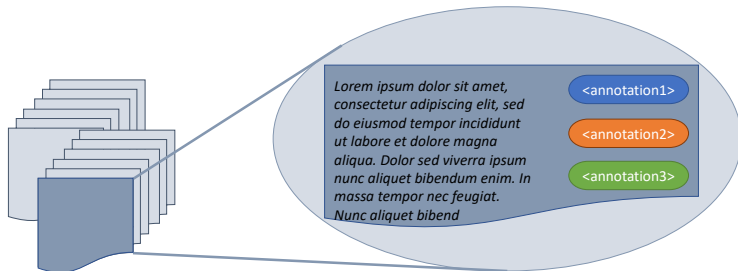
General Setting



- Corpus = set of text documents

A Corpus of Documents and Annotations

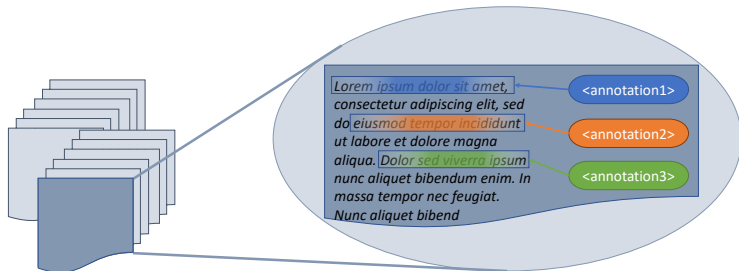
General Setting



- Corpus = set of text documents
- Each document has a set of annotations

A Corpus of Documents and Annotations

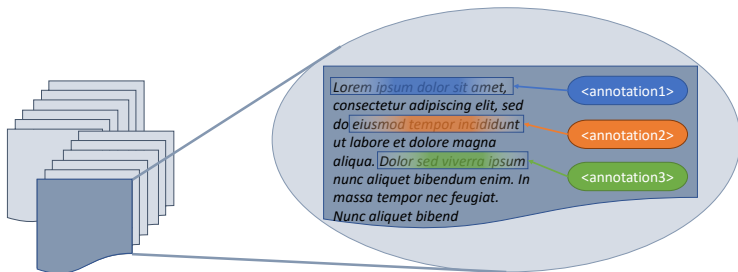
General Setting



- Corpus = set of text documents
- Each document has a set of annotations
- Each annotation is associated with words at a specific location

A Corpus of Documents and Annotations

General Setting



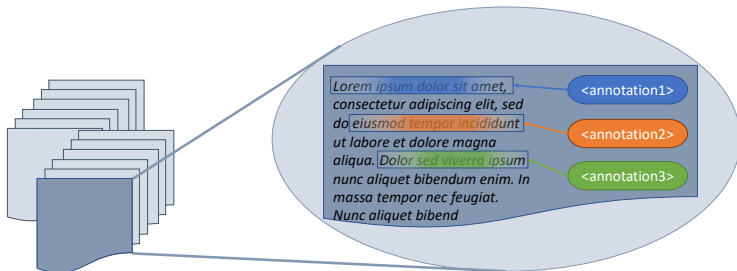
- Corpus = set of text documents
- Each document has a set of annotations
- Each annotation is associated with words at a specific location

Assumption

Annotations are relevant for a given task, i.e., reflect the context.

A Selection of Tasks in Data Linking

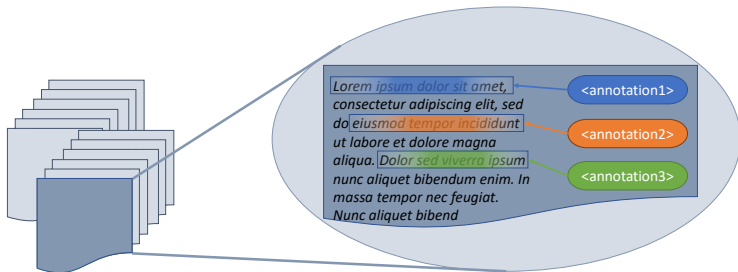
Standard Document Retrieval



In a *fixed* corpus, find documents

A Selection of Tasks in Data Linking

Standard Document Retrieval

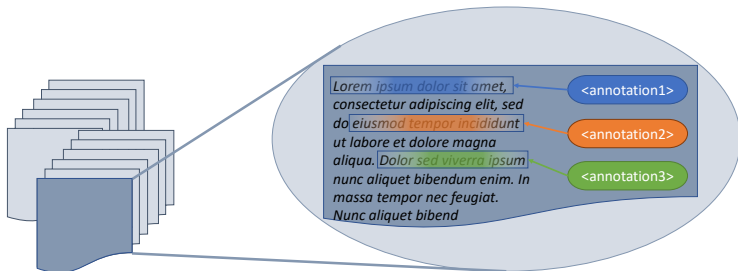


In a *fixed* corpus, find documents

- forming a cluster of documents similar in some regard,

A Selection of Tasks in Data Linking

Standard Document Retrieval

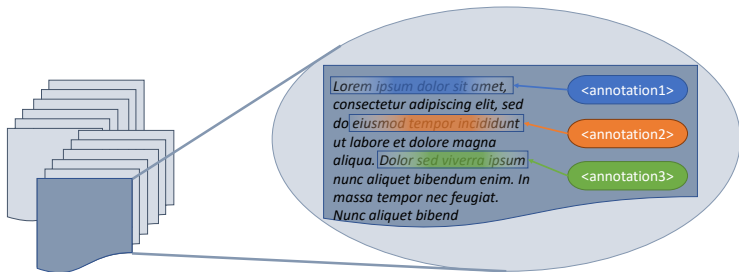


In a *fixed* corpus, find documents

- forming a cluster of documents similar in some regard,
- that fit some search terms, or

A Selection of Tasks in Data Linking

Standard Document Retrieval

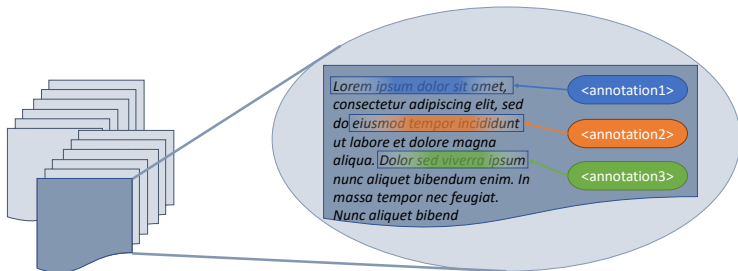


In a *fixed* corpus, find documents

- forming a cluster of documents similar in some regard,
- that fit some search terms, or
- are similar to a given document.

A Selection of Tasks in Data Linking

Standard Document Retrieval



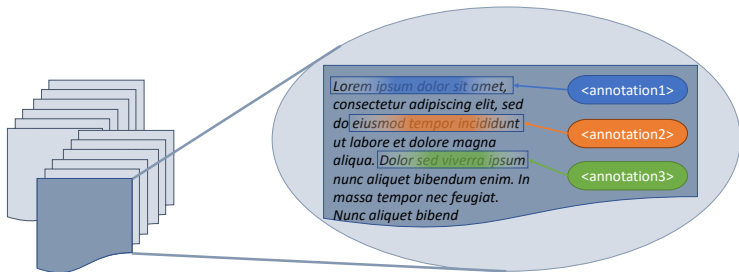
In a *fixed* corpus, find documents

- forming a cluster of documents similar in some regard,
- that fit some search terms, or
- are similar to a given document.

Solve the task using

A Selection of Tasks in Data Linking

Standard Document Retrieval



In a *fixed* corpus, find documents

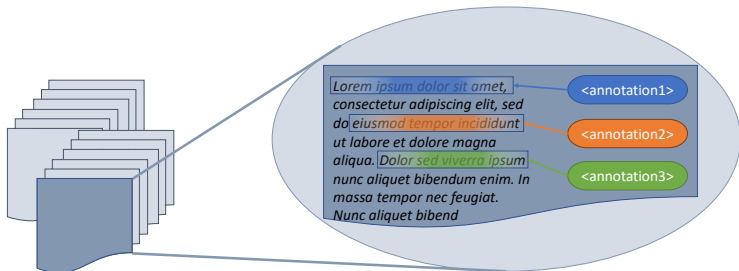
- forming a cluster of documents similar in some regard,
- that fit some search terms, or
- are similar to a given document.

Solve the task using

- words, e.g., tf.idf,

A Selection of Tasks in Data Linking

Standard Document Retrieval



In a *fixed* corpus, find documents

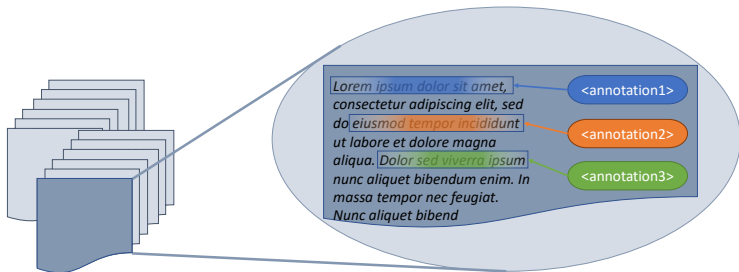
- forming a cluster of documents similar in some regard,
- that fit some search terms, or
- are similar to a given document.

Solve the task using

- words, e.g., tf.idf,
- topics, e.g., LDA, or

A Selection of Tasks in Data Linking

Standard Document Retrieval



In a *fixed* corpus, find documents

- forming a cluster of documents similar in some regard,
- that fit some search terms, or
- are similar to a given document.

Solve the task using

- words, e.g., tf.idf,
- topics, e.g., LDA, or
- annotations (depends on type of annotations), e.g., entity matching.

A Selection of Tasks in Data Linking

Corpus Extension as a Formalisation of Recommending Documents



*Lorem ipsum dolor sit amet,
consectetur adipiscing elit, sed
do eiusmod tempor incididunt
ut labore et dolore magna
aliqua. Dolor sed viverra ipsum
nunc aliquet bibendum enim. In
massa tempor nec feugiat.
Nunc aliquet bibend*

Extend a corpus with a **new document**

A Selection of Tasks in Data Linking

Corpus Extension as a Formalisation of Recommending Documents



*Lorem ipsum dolor sit amet,
consectetur adipiscing elit, sed
do eiusmod tempor incididunt
ut labore et dolore magna
aliqua. Dolor sed viverra ipsum
nunc aliquet bibendum enim. In
massa tempor nec feugiat.
Nunc aliquet bibend*

Extend a corpus with a **new document**
only if the **document**

*provides additional data relevant for a
given task, i.e., adds value in a given
context.*

A Selection of Tasks in Data Linking

Corpus Extension as a Formalisation of Recommending Documents



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Extend a corpus with a **new document** only if the **document**

provides additional data relevant for a given task, i.e., adds value in a given context.

→ **Corpus enrichment**

A Selection of Tasks in Data Linking

Corpus Extension as a Formalisation of Recommending Documents



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Extend a corpus with a **new document** only if the **document**

provides additional data relevant for a given task, i.e., adds value in a given context.

→ **Corpus enrichment**

Make decision based on

- words, BUT: not context-specific

A Selection of Tasks in Data Linking

Corpus Extension as a Formalisation of Recommending Documents



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Extend a corpus with a **new document** only if the **document**

provides additional data relevant for a given task, i.e., adds value in a given context.

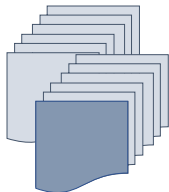
→ **Corpus enrichment**

Make decision based on

- words, BUT: not context-specific
- topics, BUT: possibly inconclusive

A Selection of Tasks in Data Linking

Corpus Extension as a Formalisation of Recommending Documents



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Extend a corpus with a **new document** only if the **document**

provides additional data relevant for a given task, i.e., adds value in a given context.

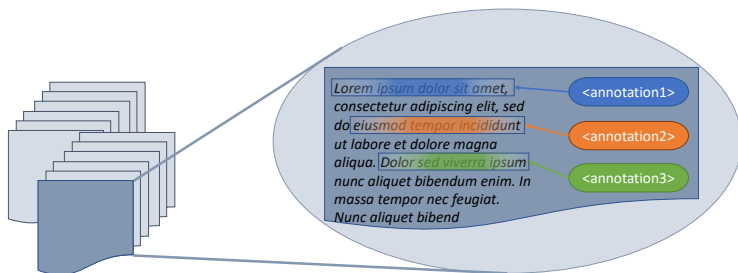
→ **Corpus enrichment**

Make decision based on

- words, BUT: not context-specific
- topics, BUT: possibly inconclusive
- *annotations?*

Annotations for Corpus Enrichment

Foundation

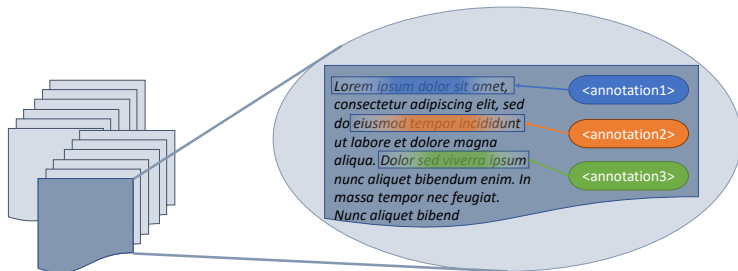


Assumption

Annotations are relevant for a given task, i.e., reflect the context.

Annotations for Corpus Enrichment

Foundation



Assumption

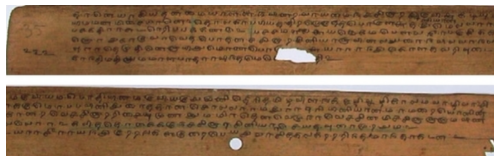
Annotations are relevant for a given task, i.e., reflect the context.

Proposition

Annotations generate the words in a document.

Understanding Written Artefacts

Foundation



*pāra+ tolkāppiyamum pattupāṭṭum kaliyum
āra+ kuṟuntokaiyuḷ aiññāṅkam – cāra+
tīru+ taku mā muṇi cey cintāmaṇiyum
virutti nacciṅārkkīyamē.*

On the weighty *Tolkāppiyam* and the *Pattuppāṭṭu* and *Kali*
and on five [times] four in the ornamental *Kuṟuntokai* and on the
essential *Cintāmaṇi* made by the brilliant great sage (Tirutakkatēvar)
[are] the elaborate commentaries [attributed] to Nacciṅārkkīyār.

Assumption

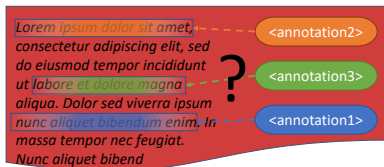
Annotations are relevant for a given task, i.e., reflect the context.

Proposition

Annotations generate the words in a document.

Annotations for Corpus Enrichment

Foundation

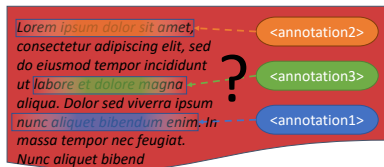


Proposition

Annotations generate the words in a document.

Annotations for Corpus Enrichment

Foundation



Proposition

Annotations generate the words in a document.

Question

How much of the document can we generate with high probability given the annotations in the corpus?

Annotations for Corpus Enrichment

How does this question help to decide “To extend or not to extend?”?



Lorem ipsum dolor sit amet, ———— <annotation2>
consectetur adipiscing elit, sed
do eiusmod tempor incididunt
ut labore et dolore magna ———— <annotation3>
aliqua. Dolor sed viverra ipsum
nunc aliquet bibendum enim, in ———— <annotation1>
massa tempor nec feugiat.
Nunc aliquet bibend

Decision

Based on answer to how much is generated with high probability:
decide extension (IN/OUT)

Annotations for Corpus Enrichment

How does this question help to decide “To extend or not to extend?”?



Lorem ipsum dolor sit amet, ———— <annotation2>
consectetur adipiscing elit, sed
do eiusmod tempor incididunt
ut labore et dolore magna ———— <annotation3>
aliqua. Dolor sed viverra ipsum
nunc aliquet bibendum enim. ———— <annotation1>
In
massa tempor nec feugiat.
Nunc aliquet bibend

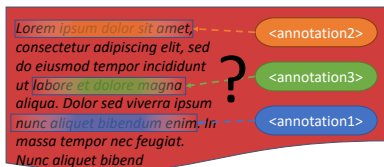
Decision

Based on answer to how much is generated with high probability:
decide extension (IN/OUT)

- Generate large part with high probability: OUT (→ known).

Annotations for Corpus Enrichment

How does this question help to decide “To extend or not to extend?”?



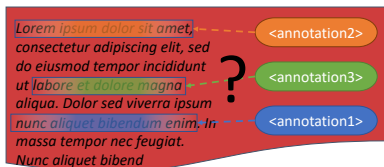
Decision

Based on answer to how much is generated with high probability:
decide extension (IN/OUT)

- Generate large part with high probability: OUT (→ known).
- Probability low: OUT (→ unrelated).

Annotations for Corpus Enrichment

How does this question help to decide “To extend or not to extend?”?



Decision

Based on answer to how much is generated with high probability:
decide extension (IN/OUT)

- Generate large part with high probability: OUT (→ known).
- Probability low: OUT (→ unrelated).
- Generate only some parts with high probability: IN (→ extension).

Annotations for Corpus Enrichment

Approach



Diagram illustrating the approach to corpus enrichment. A stack of documents is shown on the left. On the right, a red box contains a sample text with annotations:

>Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

The text is annotated with three colored boxes:

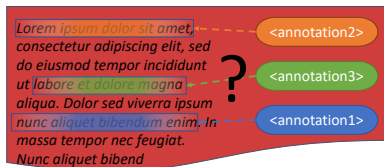
- Orange box: *consectetur adipiscing elit, sed do eiusmod tempor incididunt* (labeled <annotation2>)
- Green box: *ut labore et dolore magna aliqua* (labeled <annotation3>)
- Blue box: *nunc aliquet bibendum enim. In massa tempor nec feugiat.* (labeled <annotation1>)

A large black question mark is positioned to the right of the text, indicating a question or uncertainty about the annotations.

Annotations for Corpus Enrichment

Approach

$$\begin{matrix} & w_1 & w_2 & w_3 & \cdots & w_n \\ t_1 & \left[\begin{array}{cccccc} v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array} \right. \\ t_2 & \\ \vdots & \\ t_m & \end{matrix}$$

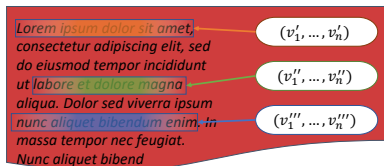


- Corpus documents (offline): build vector representation of annotations with respect to words occurring with annotations

Annotations for Corpus Enrichment

Approach

$$\begin{matrix} & w_1 & w_2 & w_3 & \cdots & w_n \\ t_1 & \left[\begin{array}{cccccc} v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array} \right. \end{matrix}$$

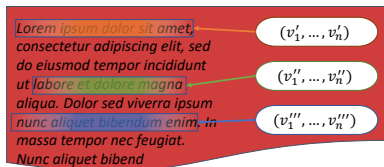


- Corpus documents (offline): build vector representation of annotations with respect to words occurring with annotations
- New document: for word chunks, build vector representation of the words occurring in the chunk

Annotations for Corpus Enrichment

Approach

$$\begin{matrix} & w_1 & w_2 & w_3 & \cdots & w_n \\ t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{matrix}$$

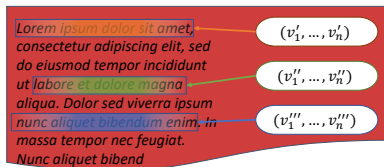
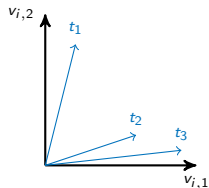


- Corpus documents (offline): build vector representation of annotations with respect to words occurring with annotations
- New document: for word chunks, build vector representation of the words occurring in the chunk
- Use cosine similarity to find annotation whose vector representation is most similar to the words of a chunk:

$$\text{sim}(A, B) = \cos(\angle A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Annotations for Corpus Enrichment

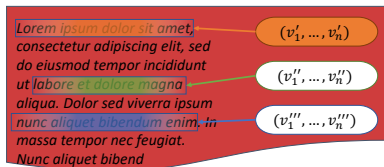
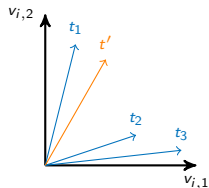
Approach



- Simplified representation of corpus annotations t_i with two words in the vocabulary

Annotations for Corpus Enrichment

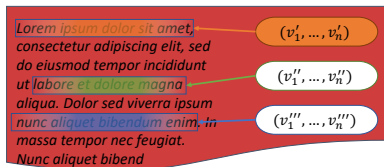
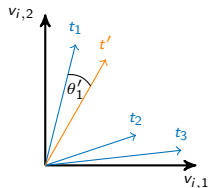
Approach



- Simplified representation of corpus annotations t_i with two words in the vocabulary
- Representation of vector representation of word chunk t'

Annotations for Corpus Enrichment

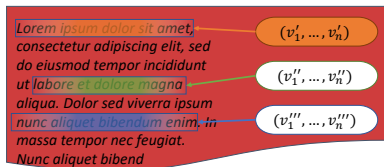
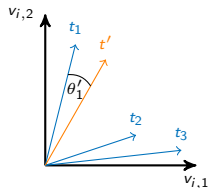
Approach



- Simplified representation of corpus annotations t_i with two words in the vocabulary
- Representation of vector representation of word chunk t'
- Angle θ'_1 between t_1 and t' smallest compared to t_2, t_3

Annotations for Corpus Enrichment

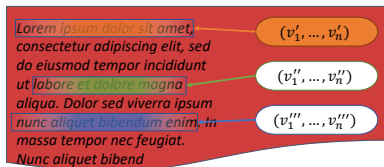
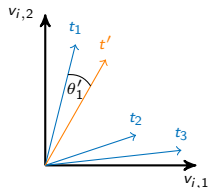
Approach



- Simplified representation of corpus annotations t_i with two words in the vocabulary
 - Representation of vector representation of word chunk t'
 - Angle θ'_1 between t_1 and t' smallest compared to t_2, t_3
- Find t_i with smallest angle for each word chunk

Annotations for Corpus Enrichment

Approach

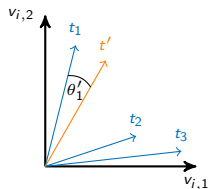


- Simplified representation of corpus annotations t_i with two words in the vocabulary
 - Representation of vector representation of word chunk t'
 - Angle θ'_1 between t_1 and t' smallest compared to t_2, t_3
- Find t_i with smallest angle for each word chunk

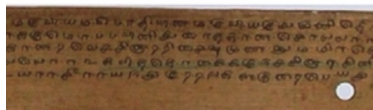
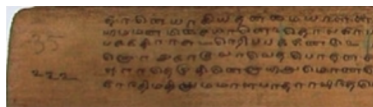
Use set of t_i 's for all word chunks t' in the new document and their similarities for decision

Understanding Written Artefacts

Possible Extension



Lorem ipsum dolor sit amet, (v'_1, \dots, v'_n)
consectetur adipiscing elit, sed
do eiusmod tempor incididunt
ut labore et dolore magna (v''_1, \dots, v''_n)
aliqua. Dolor sed viverra ipsum
nunc aliquet bibendum enim, in (v'''_1, \dots, v'''_n)
massa tempor nec feugiat.
Nunc aliquet bibend

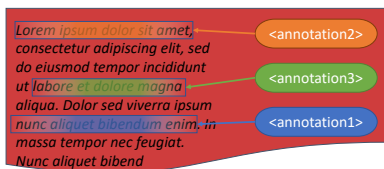


Embed further data into corpus vector representation, e.g.,

- materials
 - spectrograms
- Cosine similarity over words *and* materials ($(, \dots)$)
- Use for corpus enrichment or document retrieval

Annotations for Corpus Enrichment

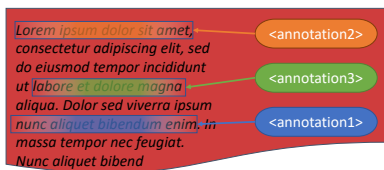
Results (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)



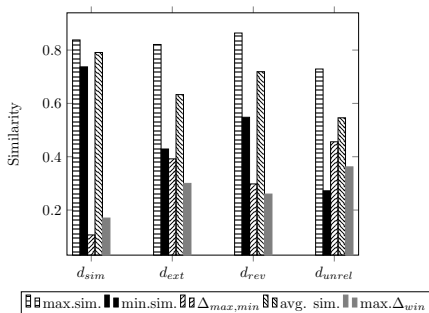
- Corpus: Wikipedia articles about cities
- New document:
 - d_{sim} : known
 - d_{ext} : extended
 - d_{rev} : revised
 - d_{unrel} : unrelated

Annotations for Corpus Enrichment

Results (from: Kuhr, B, Bender, Möller, "To Extend or Not to Extend? Context-driven Corpus Enrichment", AI 2019.)

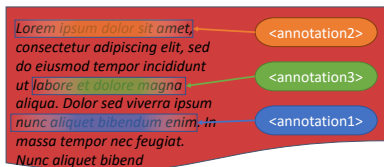


- Corpus: Wikipedia articles about cities
- New document:
 - d_{sim} : known
 - d_{ext} : extended
 - d_{rev} : revised
 - d_{unrel} : unrelated



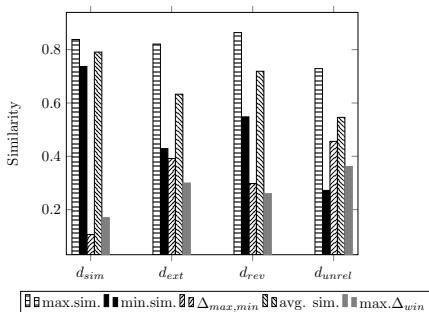
Annotations for Corpus Enrichment

Challenges (from: Kuhr, B, Bender, Möller, "To Extend or Not to Extend? Context-driven Corpus Enrichment", AI 2019.)



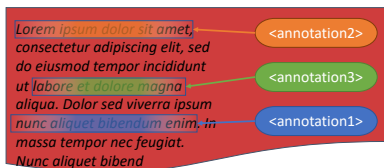
Influencing factors

- Corpus size
 - Quality of annotations
 - Indicators
- No single indicator to rule them all!
- Limited transfer between corpora!



Annotations for Corpus Enrichment

Challenges (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)

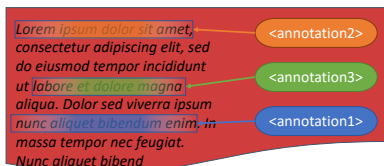
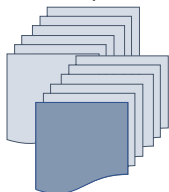


Indicator I	city corpus				president corpus			
	d_{sim}	d_{ext}	d_{rev}	d_{unrel}	d_{sim}	d_{ext}	d_{rev}	d_{unrel}
Max Sim.	+	+	+	o	+	+	+	o
Min Sim.	+	o	o	-	o	o	o	-
$\Delta_{max,min}$	-	o	-	o	-	o	-	o
Avg. Sim.	+	o	+	o	+	+	+	o
Max. Δ_{win}	-	o	-	o	-	o	-	o

“+”: $I \geq 0.7$, “-”: $I \leq 0.3$, “o”: $0.3 < I < 0.7$

Annotations for Corpus Enrichment

Challenges (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)

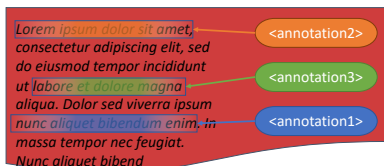


Indicator I	city corpus				president corpus			
	d_{sim}	d_{ext}	d_{rev}	d_{unrel}	d_{sim}	d_{ext}	d_{rev}	d_{unrel}
Max Sim.	+	+	+	o	+	+	+	o
Min Sim.	+	o	o	-	o	o	o	-
$\Delta_{max,min}$	-	o	-	o	-	o	-	o
Avg. Sim.	+	o	+	o	+	+	+	o
Max. Δ_{win}	-	o	-	o	-	o	-	o

“+”: $I \geq 0.7$, “-”: $I \leq 0.3$, “o”: $0.3 < I < 0.7$

Annotations for Corpus Enrichment

Challenges (from: Kuhr, B, Bender, Möller, “To Extend or Not to Extend? Context-driven Corpus Enrichment”, AI 2019.)

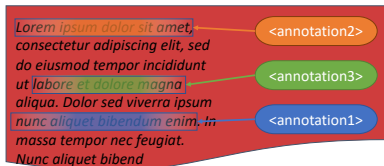


Indicator I	city corpus				president corpus			
	d_{sim}	d_{ext}	d_{rev}	d_{unrel}	d_{sim}	d_{ext}	d_{rev}	d_{unrel}
Max Sim.	+	+	+	o	+	+	+	o
Min Sim.	+	o	o	-	o	o	o	-
$\Delta_{max,min}$	-	o	-	o	-	o	-	o
Avg. Sim.	+	o	+	o	+	+	+	o
Max. Δ_{win}	-	o	-	o	-	o	-	o

“+”: $I \geq 0.7$, “-”: $I \leq 0.3$, “o”: $0.3 < I < 0.7$

Annotations for Corpus Enrichment

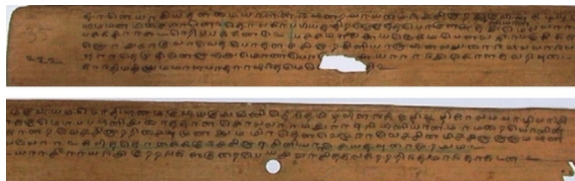
Benefits



- Enrich corpus with documents that add value
→ Recommendations
- Use similarities as guideline to unknown portions
- Use annotations as a starting point for annotating new document
- Augment annotations of corpus documents with new annotations of unknown document portions

Understanding Written Artefacts

Possible Directions



Transcript

*pāra+ tolkāppiyamum pattupāṭṭum kaliyum
āra+ kuṟutokaiyu! aiññāṅkum – cāra+
tīru+ taku mā muṇi cey cintāmaṇiyum
virutti nacciṅārkkīṇiyamē.*

Translation

On the weighty *Tolkāppiyam* and the *Pattuppāṭṭu* and *Kali*
and on five [times] four in the ornamental *Kuṟutokai* and on the
essential *Cintāmaṇi* made by the brilliant great sage (Tirutakkaṭēvar)
[are] the elaborate commentaries [attributed] to Nacciṅārkkīṇiyar.

- Parse critical editions/manuscripts (and existing annotations)
- Annotate translations with translations of referenced books (managing copyrights)
- Add annotations to manuscript without annotations using books that reference the manuscript